Some figures concerning the transliteration of the Dutch Speech Styles Corpus

Els den Os, Marian Ellens, Cor in 't Veld, and Lou Boves
SPEX, Leidschendam, The Netherlands

ABSTRACT

On the basis of orthographic transliterations of monologues, picture descriptions and read short texts of in total 127 speakers (in total more than 19 hours of speech), we present data concerning speech rate, hesitational sounds, clitic groups, and verbally deleted words (repetitions, repairs at the beginning of an utterance (false starts), and repairs later in an utterance).

INTRODUCTION

The Dutch Speech Styles Corpus was collected to investigate the voice quality of speakers of standard Dutch. The speech material was designed by R. van Bezooijen and the speech recordings were made by J. van Rie and R. van Bezooijen. The corpus contains three different speech styles: spontaneous speech (monologues), semi-spontaneous speech (picture descriptions), and read speech. The speech was always recorded in the presence of a female 'interviewer' of about 30 years old. In all three styles the speech contents refer to domestic topics, eating habits, and food.

There are 127 speakers, in three age categories: 30 speakers (17 males and 13 females) from 10 to 20 years old, 45 speakers (19 males and 26 females) from 20 to 60 years old, and 52 speakers (24 males and 28 females) between 60 and 86 years old. The total duration of speech is 19 hours and 10 minutes (4 hours and 40 minutes of monologues, 10 hours and 20 minutes of picture descriptions, and 4 hours and 10 minutes of read speech). The total number of *word forms* that were transcribed in the corpus amounts to about 118,000. There are about 6,300 *different* word forms.

The whole corpus has been transliterated. Among other things, these transliterations offer the possibility to study disfluencies in several speech styles. The goal of this paper is to compare disfluencies in two types of spontaneous speech and read speech. By disfluencies we mean in this paper specifically hesitational sounds (filled pauses like 'uh'), and verbally deleted words, i.e. words spoken, but superseded by subsequent speech.

TRANSLITERATION

The transliteration of the corpus is a word level transcription of what the speakers said. The standard spelling of Dutch is used. This necessarily implies a compromise between the sounds heard and what has to be written down.

Because it could be expected that some reduced forms of words (mostly containing schwa's) would occur more often than the full forms, it was allowed to write these reduced forms down in a sometimes non-standard way; see [1] for a more detailed description of the transliteration of the corpus.

The Speech Styles Corpus has been labelled at the utterance level (i.e. a time stamp between utterances is provided to allow access to the speech files). The notion of what constitutes an utterance in spontaneous speech is necessarily an arbitrary one. An utterance was defined as a number of words being semantically consistent and containing at least a subject and a verb. In addition, this string had to be preceded and followed by a clear acoustic pause.

Clitization

Clitization which resulted in syllable deletion was indicated and for this we could not always use existing spellings. However, we decided to mark these forms, since we wanted to know how often these forms occur in spontaneous Dutch speech. These forms are of significance in relation to automatic segmentation programmes and training of speech recognizers.

MARKINGS

Next to the orthographic transcriptions, conventions were used to indicate all audible events that occur during speaking. These conventions consist of different kinds of brackets with or without additional information, see also [2] and [3] for comparable markings used in the ATIS and Switchboard corpus. Only those conventions will be presented below that will be discussed in the following.

Hesitational sounds

A distinction was made between hesitational sounds which were uttered in isolation and those which were uttered connected to the preceding word. There were four types of hesitational sounds pauses: [uh], [um], [mm], and [naa]. For a hesitational sound to be isolated, a silent pause has to occur before and after it.

Words spoken by the interviewer

Words spoken by the interviewer are indicated with curly brackets { }. The interviewer interfered most in the monologues (3700 words) opposed to 1021 words in the picture descriptions and 1 word in the read texts. Speech rate per style was calculated on the basis of the total number of words (those by the speakers as well as those by the interviewer). This procedure had to be followed, since we could only use the total duration per style per speaker to calculate speech rate. Since the interviewer was the same person most of the time, we think the used procedure is justified.

Verbally deleted words

Words verbally deleted by the subject are enclosed in angle brackets. Verbal deletions are words spoken by the speaker but which are superseded by subsequent speech. This can occur explicitly (<at> <the> <grocery> <I> <mean> at the bakery.....) or implicitly (<at> <the> <grocery> at the bakery...). Both can occur at the beginning of an utterance (false start) or later in an utterance. Verbally deleted words can be literally repeated or can be repaired. Word fragments are also indicated by angle brackets (<ba> bakery).

After the transliterations were completed, the utterances containing angle brackets were selected, and a classification was made in different types of verbally deleted words:

(1) *repetitions*: literally repeated words, word groups or word fragments. They do *not* occur at the beginning of an utterance.

(2) *false starts*: the speaker starts an utterance producing a word, word group or word fragment, but he/she decides to start all over again. The beginning of an utterance was defined as the first two words.

(3) *repairs later in the utterance*: the speakers interrupts a word, word group or word fragment and continues the utterance in a different way.

SPEECH RATE

The overall speech rate measured in words per minute was somewhat higher for reading texts (156 words per minute) than for monologues and picture descriptions (136 and 129 words per minute respectively). No difference was observed between the speaking rates of male and female speakers. Nor was any great difference observed between the various age categories, although in the

monologues the young speakers spoke somewhat more slowly (123 words per minute) than the older speakers (143 words per minute). It must be noted that in this calculation the time spent in pausing was included.

CLITIZATION

The total number of clitic forms that resulted in syllable deletion amounted to about 450. This is only 0.4% of the total number of words in the corpus. Especially the forms including the personal pronoun *ik* 'I', the verb form *is* 'is', and the personal pronoun *het* 'it' involve syllable deletion.

FILLED PAUSES

Filled pauses only occurred in the monologues and in the picture descriptions. There were no differences in the number of filled pauses between these two styles.

Related to the total number of words produced per speech style by male or female speakers of one of the three age categories, the percentage of filled pauses is between 2.6% (picture descriptions by male speakers of 20 to 60 years of age) and 7.5% (picture descriptions by male speakers under 20 years of age). The younger speakers produced more filled pauses (on average 7%) than the older ones (on average 5%). In both styles the relative number of connected filled pauses was about the same as the relative number of isolated filled pauses (2.8% and 2.3%, respectively). In Dutch, speakers often connect pauses to function words, e.g. *en[uh]* 'and[uh]'. We only observed very few instances in which the filled pause was connected to a content word.

VERBALLY DELETED WORDS

In the following, we present the percentage of words that are themselves verbally deleted or are involved in a verbal deletion. The data are given for male and female speakers, and for the age groups separately. We only present data for the monologues and picture descriptions, because verbally deleted words did almost not occur in the read texts.

Repetitions

In table 1 it can be observed that there are no clear differences between the age groups, sex, or styles for the percentage of repetitions. The percentages range from 0.4 to 1.2.

Table 1: Percentage of repetitions for Female (F) and Male (M) speakers for the three age categories (1=<20, 2= between 20 and 60, and 3=>60 years of age), for monologues and picture descriptions.

	Monologues	Picture description
F1	0.8%	1.0%
F2	0.8%	0.7%
F3	1.0%	0.5%
M1	0.6%	0.9%
M2	1.1%	0.4%
M3	1.2%	0.8%

False starts

In table 2, the percentages of false starts are given. It can be seen that the younger speakers produce more false starts than the older ones in the picture descriptions. The male and female speakers between 20 and 60 years of age produce relatively few false starts. The percentages range from 0.4 to 2.1.

Verbally deleted words later in the utterance

In table 3 the percentages of verbally deleted words later in the utterance are given. It can be observed that the older male speakers produce most verbally deleted words (2.9%) in the monologues and that the young male speakers produce most verbally deleted words in the picture descriptions (2.2%).

Table 2: Percentage of false starts for Female (F) and Male (M) speakers for the three age categories (1=<20, 2= between 20 and 60, and 3=>60 years of age), for monologues and picture descriptions.

	Monologues	Picture description
F1	0.6%	2.1%
F2	0.5%	0.6%
F3	0.9%	0.8%
M1	1.3%	1.3%
M2	0.4%	0.4%
M3	0.8%	0.8%

Table 3: Percentage of verbally deleted words later in the utterance for Female (F) and Male (M) speakers for the three age categories (1=<20, 2= between 20 and 60, and 3=>60 years of age), for monologues and picture descriptions.

	Monologues	Picture description
F1	1.1%	1.3%
F2	1.1%	1.6%
F3	1.4%	1.6%
M1	2.4%	2.2%
M2	1.3%	0.9%
M3	2.9%	1.8%

DISCUSSION AND CONCLUSION

Here we will concentrate on the disfluencies in spontaneous and semi-spontaneous speech, since it turned out that disfluencies in the read text were almost not present. This is due to the fact that the sentences in the texts were short and simple. We all know that spontaneous speech is not fluent: speakers produce many hesitational sounds, mispronunciations, and verbal deletions. As far as we know, the number of these disfluencies have never been addressed on the basis of a large number of speakers of different age groups.

Our counts show that hesitational sounds occur on average about once every 20 words. Verbally deleted words (repetitions, false starts, and deleted words later in the utterance taken together) occur on average 3 times in every hundred words. The group of male and female speakers between the ages 20 and 60 years produced fewer verbally deleted words than the younger and older group. Especially in the picture descriptions, the younger speakers produced relatively many verbal deletions.

It must be remarked here, that most verbal deletions were repaired *implicitly*. There were very few instances of explicitly repaired deletions, like <dog> <I> <mean> cat. Furthermore, it must be noted that the disfluencies mentioned above were actually repaired; there are only few instances of disfluencies which were *not* repaired by the speaker.

Most verbal deletions occurred after a word was finished (77% of all verbally deleted words in the monologues and picture descriptions together). Verbal deletions after word fragments occurred less frequently. In the corpus, we observed very few instances of silent pauses within words (43 times). In most of these cases these pauses occur between the two parts of a compound. In addition, hesitational sounds almost never occurred within words. From this, and the fact that most verbally deleted words have been completed by the speakers, we may conclude that words are preferably articulated as a whole.

REFERENCES

[1] Den Os, E.A. (1994), "Transliteration of the Dutch Speech Styles Corpus", *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 18, 87-94.

[2] Hirschmann, L. (1992), "Multi-site data collection for spoken language corpus", *Proceedings ICSLP'92*, 2, 903-906.

[3] Godfrey, J.J., Holliman, E.C. & McDaniel, J. (1992) "Switchboard: Telephone Speech corpus for Research and Development", *Proceedings ICASSP*, 1, I-517 - I-520.